European Journal of Clinical Nutrition (2012) 66, 569–576

0954-3007/12
www.nature.com/ejcn

ORIGINAL ARTICLE

The accuracy of the Goldberg method for classifying misreporters of energy intake on a food frequency questionnaire and 24-h recalls: comparison with doubly labeled water

JA Tooze[1], SM Krebs-Smith[2], RP Troiano[2] and AF Subar[2]

[1]*Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, NC, USA and* [2]*Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD, USA*

Background/Objectives: Adults often misreport dietary intake; the magnitude varies by the methods used to assess diet and classify participants. The objective was to quantify the accuracy of the Goldberg method for categorizing misreporters on a food frequency questionnaire (FFQ) and two 24-h recalls (24HRs).
Subjects/Methods: We compared the Goldberg method, which uses an equation to predict total energy expenditure (TEE), with a criterion method that uses doubly labeled water (DLW), in a study of 451 men and women. Underreporting was classified using recommended cut points and calculated values. Sensitivity and specificity, positive predictive value (PPV) and negative predictive value and the area under the receiver operating characteristic curve (AUC) were calculated. Predictive models of underreporting were contrasted for the Goldberg and DLW methods.
Results: AUCs were 0.974 and 0.972 on the FFQ, and 0.961 and 0.938 on the 24HR for men and women, respectively. The sensitivity of the Goldberg method was higher for the FFQ (92%) than the 24HR (50%); specificity was higher for the 24HR (99%) than the FFQ (88%); PPV was high for the 24HR (92%) and FFQ (88%). Simulation studies indicate attenuation in odds ratio estimates and reduction of power in predictive models.
Conclusions: Although use of the Goldberg method may lead to bias and reduction in power in predictive models of underreporting, the method has high predictive value for both the FFQ and the 24HR. Thus, in the absence of objective measures of TEE or physical activity, the Goldberg method is a reasonable approach to characterize underreporting.
European Journal of Clinical Nutrition (2012) **66,** 569–576; doi:10.1038/ejcn.2011.198; published online 30 November 2011

Keywords: diet; diet surveys; energy intake; statistical bias; questionnaires/standards; research design

Introduction

It is well accepted that adults misreport their dietary intake on self-administered tools, most often in the direction of underreporting energy intake. In many studies of underreporting, participants are classified as underreporters (URs) or acceptable reporters (ARs), the prevalence of underreporting is estimated, and personal characteristics are related to reporting status (Macdiarmid and Blundell, 1998; Hill and Davies, 2001; Livingstone and Black, 2003). Other studies have proposed excluding URs from analyses to reduce the effects of measurement error on relationships between diet and obesity or other health outcomes; exclusion of URs often leads to different conclusions than when they are included (Drummond *et al*., 1998; Huang *et al*., 2005).

The Goldberg approach is commonly used to identify misreporters (Black *et al*., 1991; Goldberg *et al*., 1991; Black, 2000a). However, because of the assumptions and formulas used to estimate total energy expenditure (TEE), it may be prone to misclassification, which could lead to bias in studies using this method. One way to test this is to use an unbiased estimate of TEE, such as that estimated from doubly labeled

Correspondence: Dr JA Tooze, Department of Biostatistical Sciences, Wake Forest School of Medicine, Medical Center Boulevard, Winston-Salem, NC 27157, USA.
E-mail: jtooze@wakehealth.edu
Received 25 March 2011; revised 24 October 2011; accepted 25 October 2011; published online 30 November 2011

water (TEE_{DLW}), to examine how well the Goldberg method classifies URs. Two studies have used TEE_{DLW} to examine the sensitivity and specificity of the Goldberg method for categorizing misreported reported energy intake (rEI; Black, 2000b; Livingstone *et al*., 2003). Both studies reported that ~50% of the participants categorized as URs using TEE_{DLW} (UR_{DLW}) were also categorized as URs by the Goldberg method (UR_{GB}). More than 98% of the participants identified by DLW as ARs (ARD_{DLW}) also were identified as AR by the Goldberg method (AR_{GB}). However, these analyses were based on small studies (Livingstone *et al*., 2003) or a pooled analysis of multiple small studies treated as a large study (Black, 2000b). Both studies used food diaries to estimate rEI; we are not aware of studies using other methods of dietary assessment.

This paper uses a large DLW sample from the Observing Protein and Energy Nutrition (OPEN) Study to compare the Goldberg method for categorizing misreporting to estimates using TEE_{DLW}. Two different dietary assessment instruments are used to estimate rEI, a food frequency questionnaire (FFQ) and two 24-h recalls (24HRs). The purpose of this paper is to compare the classification of UR using the Goldberg method with UR classified using TEE_{DLW}.

Subjects and methods

Study population
The OPEN Study is described in detail elsewhere (Subar *et al*., 2003). The primary goal of the study was to describe the measurement error structure of an FFQ and 24HRs. Participants were 484 men and women, aged 40–69 years, recruited from a random sample of 5000 households in the Washington, DC metropolitan area. In all, 58% of eligible participants agreed to participate in the study; only two participants dropped out during the course of the study. The National Cancer Institute's (NCI) Special Studies Institutional Review Board approved the protocol. Participants completed three clinic visits over a period of ~3 months between September 1999 and March 2000.

Energy intake
Participants completed an FFQ and a 24HR twice, ~3 months apart. The FFQ was the NCI Diet History Questionnaire (http://riskfactor.cancer.gov/DHQ), which was validated in previous research for this population (Subar *et al*., 2001). In this analysis, rEIs from the first FFQ were used. Trained interviewers administered the 24HR using a standardized five-pass method developed by the US Department of Agriculture (Conway *et al*., 2003, 2004; Moshfegh *et al*., 2008). The 24HR data were analyzed using the Food Intake Analysis System (version 3.99). The average of the two 24HRs was used because it is a commonly used, albeit naive practice to decrease within-person variation in the estimated usual intake.

Energy expenditure
The DLW measurement for the OPEN Study is described in detail elsewhere (Trabulsi *et al*., 2003). A five-urine specimen protocol was used (Schoeller, 1992). TEE_{DLW} was calculated according to Racette *et al*. (1994) using the modified Weir equation with a respiratory quotient of 0.86. A total of 33 TEE_{DLW} measures were excluded for the following reasons: unacceptable internal agreement ($n=2$), failure to isotopically equilibrate on dosing day ($n=10$), isotopic dilution space ratios outside the range of 1.00–1.08 ($n=6$), lack of tracer in the final urine specimen due to high water turnover ($n=5$) or missing specimens ($n=10$), resulting in 451 participants who were used in this analysis. Twenty-five participants were dosed with DLW a second time ~2 weeks after the first to obtain within-person variation of TEE_{DLW}. Weight was measured at all visits under standardized conditions. Height was measured at visit 1. Basal metabolic rate (BMR) was calculated from weight, height and age using the equation developed by Schofield (1985) for adults.

Additional measures
At visit 1, participants completed the Physical Activity Questionnaire from the National Health and Nutrition Examination Survey 1999–2000. At visit 2, ~2 weeks later, participants completed a health questionnaire that contained the Fear of Negative Evaluation Scale (Leary, 1983) and questions regarding Stunkard–Sorenson body silhouettes (Stunkard *et al*., 1982). At visit 3 (~3 months after visit 1), participants completed the Three-Factor Eating Questionnaire (Stunkard and Messick, 1985), the Marlowe–Crowne Social Desirability Questionnaire (Crowne and Marlowe, 1960; Strahan and Gerbasi, 1972; Fischer and Fick, 1993) and questions about dieting/weight loss.

Classification of misreporters
In the DLW method and the Goldberg method, participants are classified as URs, ARs or overreporters using the ratio of rEI to TEE. In the Goldberg method, TEE_{GB} is calculated from the product of BMR and physical activity level (PAL). A constant value is assumed for PAL, and therefore the ratio of rEI:TEE may be expressed in terms of multiples of rEI to BMR. Because of skewness observed in the distribution of energy intake, the natural log transformation of the ratio is used in both methods. In both the DLW method and the Goldberg method, a 95% confidence interval is created about the log of the ratio, and individuals who fall outside of the confidence interval are classified as URs or overreporters.

For the Goldberg method, values for variation in rEI, BMR and PAL as suggested by Black (2000b) were applied to classify misreporting. PAL was assumed to be 1.55. In secondary analyses, we classified URs using a different assumption for variability on the FFQ; in particular, we used the coefficient of variation from the OPEN Study to estimate

within-person variation for 1 day of measurement (Supplementary Material).

Statistical analysis

Sensitivity and specificity analyses. Because DLW is an objective biomarker of TEE, and therefore a marker of energy intake under energy balance, the classification of reporting status using rEI:TEE_{DLW} was the 'gold standard' in our analyses. Owing to the small numbers of participants classified as overreporters, this group was excluded from the sensitivity and specificity analyses. Sensitivity was calculated as the proportion of UR_{GB} among UR_{DLW}. Specificity was calculated as the proportion of AR_{GB} among AR_{DLW}. Positive predictive value (PPV), the probability of being an UR if classified as one by the Goldberg method, and Negative predictive value (NPV), the probability of being an AR if classified as one by the Goldberg method, were calculated. We also used the area under the receiving operator characteristic curve (AUC) to quantify the classification accuracy of the Goldberg method. AUC over 0.9 indicates outstanding discrimination (Hosmer and Lemeshow, 2000).

Cut point and TEE analyses. The differences between the Goldberg method and the DLW method for classifying URs are due to the: estimate of TEE from the Goldberg formula or DLW, and cut points used (Figures 1 and 2). If $TEE_{GB} = TEE_{DLW}$, the ratio of rEI:TEE is equivalent, and the two methods agree. Even if TEE_{GB} differs from TEE_{DLW}, the two methods will provide the same classification if the participants are below (or above) both of the cut points for the two methods. We estimated whether differences between the methods were due to differences in the cut point (participants were between the cut points for the two methods) or due to TEE (TEE would lead to discrepancies even with the same cut points). We compared TEE_{GB} with TEE_{DLW} using a Wilcoxon signed-rank test and by calculating correlation (Supplementary Material).

Analysis of implications of using the Goldberg method compared with the DLW method. To study the implications of using TEE_{GB} to classify URs in studies of characteristics of URs, we modeled the probability of being an UR_{GB} using variables previously identified as statistically significant predictors of UR_{DLW} in the OPEN Study (Tooze *et al*., 2004). The variables include: education (men, 24HR), body mass index (BMI; men, FFQ; women and men, 24HR), percent of energy from fat (women, FFQ and 24HR), number of eating occasions (men, FFQ and 24HR), variability in number of meals (women, 24HR), whether the participant has ever lost 10 lbs or more (women, FFQ), times dieted (men, 24HR), fear of negative evaluation (women, FFQ), activity level compared with others (men, FFQ), usual activity (women,

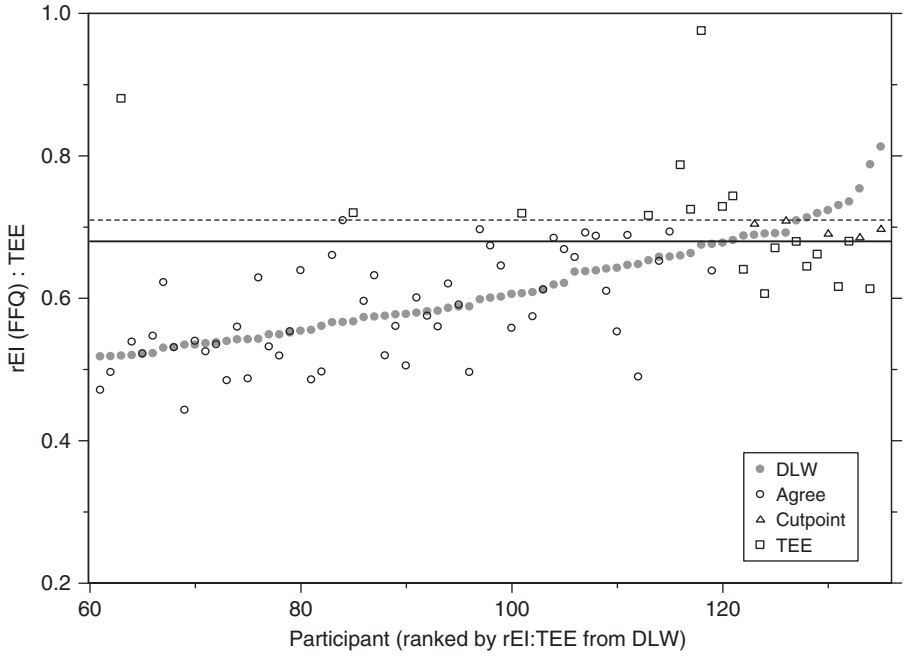

Figure 1 Men: ratio of rEI on a FFQ to TEE, as estimated by DLW (illustrated with filled circles) or the Goldberg method (circles, triangles and squares) by participant, ranked by ratio from DLW value. Only the participants classified as URs by either method (FFQ: $n = 136$) are shown in the figure; for clarity the first 60 men (who showed agreement) are excluded from the plot. Open circles indicate that the Goldberg method classification agrees with DLW classification; triangles indicate that the difference between the two methods is due to differences in the cut points; and squares indicate that the differences are due to estimation of TEE. The dashed line represents the cut point from the Goldberg method (0.71), and the solid line represents the cut point from DLW (0.68).

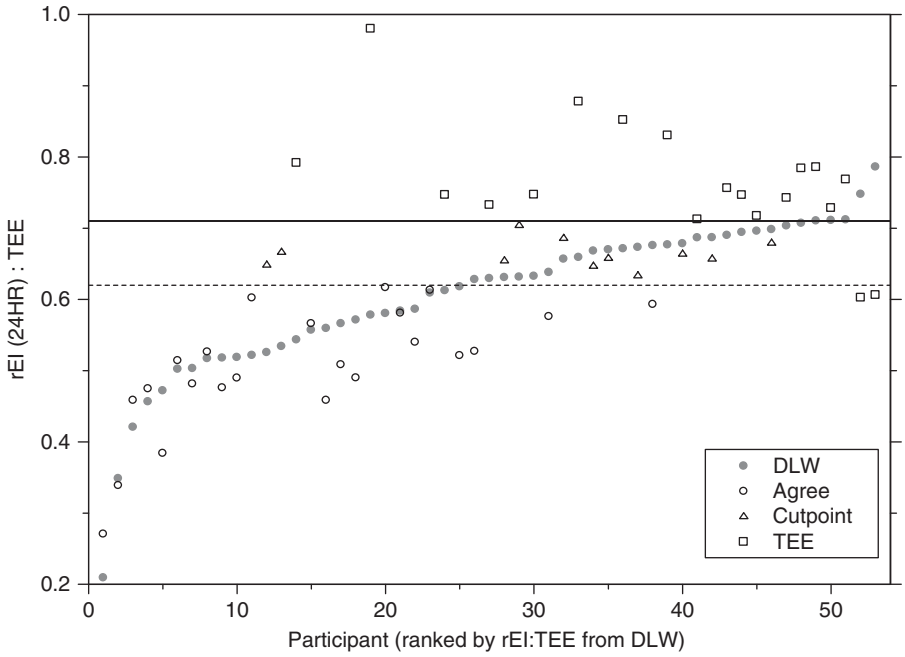

Figure 2 Men: ratio of rEI on the average of two 24HRs to TEE, as estimated by DLW (illustrated with filled circles) or the Goldberg method (circles, triangles and squares) by participant, ranked by ratio from DLW value. Only the participants classified as URs by either method (24HR: $n = 53$) are shown in the figure. Open circles indicate that the Goldberg method classification agrees with DLW classification; triangles indicate that the difference between the two methods is due to differences in the cut points; and squares indicate that the differences are due to estimation of TEE. The dashed line represents the cut point from the Goldberg method (0.62), and the solid line represents the cut point from DLW (0.71).

24HR), social desirability (women and men, 24HR) and restraint (men, 24HR).

We also did a simulation study to quantify the effects of varying sensitivities and specificities in this type of model. BMI (mean = 27.9, s.d. = 5.3) was simulated based on the participants in the OPEN Study for 300 data sets with 500 individuals each. True UR status was simulated from BMI, with 49% probability of classification as a true UR, and a 35% increase in the odds of being an UR for each 5 kg/m^2 increase in BMI, based on the observed relationship in the OPEN Study for the FFQ (Tooze *et al*., 2004). Finally, the UR_{GB} status was simulated for five different combinations of sensitivity and specificity, and the relationship between UR_{GB} was modeled in a logistic regression for each data set. All analyses were performed using SAS software (version 9; Cary, NC, USA).

Results

As reported previously (Subar *et al*., 2003), using TEE_{DLW} 21% of men and 22% of women were categorized as URs on the 24HR, and 50% of men and 49% of women were categorized as URs on the FFQ. Using TEE_{GB} and the standard cut points recommended by Black (2000a), 10% of men and 13% of women were categorized as URs on the 24HR, and 52% of men and 51% of women were categorized as URs on the FFQ. The AUC analysis indicated outstanding discrimination for men and women for both instruments using UR_{GB}. The AUCs were 0.974 and 0.972 for the FFQ, and 0.961 and 0.938 for the 24HR, for women and men, respectively.

For the FFQ, sensitivity of the Goldberg method for identifying URs was 92.6% for men and 92.1% for women; specificity was 87.6% for both men and women (Table 1). The PPV was 88% and NPV was 92% for both men and women. When we assumed that the rEI was based on one measure and not infinite as the Goldberg method commonly assumes (Black, 2000b), and used the estimate of within-person variation from the OPEN Study in the formula, sensitivity was lower (71.9% for men, 62.4% for women), and specificity increased (100% for men, 99% for women). Sensitivity for the 24HR was 45.1% for men and 54.3% for women; specificity was 98.9% for men and 95.5% for women. The PPV was 92% for men and women; the NPV was 86% for men and 88% for women.

For sensitivity for the FFQ, in both men and women 100% (9/9 for men and 8/8 for women) of misclassification was due to differences in the estimate of TEE (Figure 1). For specificity on the FFQ, for men 64% (9/14) of misclassification was due to differences in the estimate of TEE, and 36% (5/14) was due to differences in the cut points; for women 25% (3/12) was due to TEE, and 75% (9/12) due to the cut points. For sensitivity for the 24HR, in men 61% (17/28) of misclassification was due to differences in the estimate of TEE, and

Table 1 Sensitivity and specificity of Goldberg method for the FFQ and 24HR in the Observing Protein and Energy Nutrition Study

		Classification by TEE_{DLW} (n = 451)											
				UR		*AR*			*OR*				
		Cut points[a]				*Classification by TEE_{GB}*					Sensitivity and specificity		
Instrument	*Sex*	*Lower*	*Upper*	*UR* (n)	*AR* (n)	*UR* (n)	*AR* (n)	*OR* (n)	*AR* (n)	*OR* (n)	*Misclassified (%)*	*Sensitivity (%)*	*Specificity (%)*
FFQ	M	1.10	2.19	112	9	14	99	4	0	6	11.1	92.6	87.6
	F	1.10	2.19	93	8	12	85	4	0	4	11.6	92.1	87.6
24HR	M	0.96	2.49	23	28	2	188	0	2	2	13.1	45.1	98.9
	F	0.96	2.49	25	21	2	156	0	0	2	11.1	54.3	98.7

Abbreviations: AR, acceptable reporter; F, female; FFQ, food frequency questionnaire; M, male; OR, overreporter; TEE_{DLW}, total energy expenditure as estimated by doubly labeled water; TEE_{GB}, total energy expenditure as estimated by Goldberg method; UR, underreporter; 24HR, 24-h recall.
[a]Cut point is ratio of reported energy intake divided by basal metabolic rate.

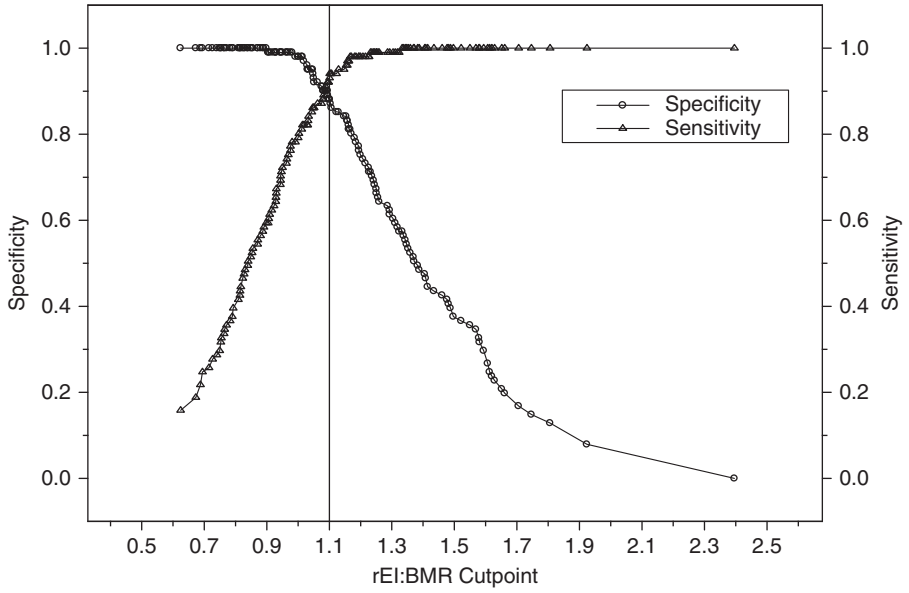

Figure 3 Plot of sensitivity and specificity by rEI:BMR cut point for the FFQ for women. The vertical line indicates the cut point from the standard Goldberg method using the values suggested by Black (2000a) (cut point = 1.10).

39% (11/28) due to differences in the cut point (Figure 2); for women 76% (16/21) was misclassified due to differences in the estimate of TEE, and 24% (5/21) was misclassified due to differences in the cut point. For specificity on the 24HR, 100% (2/2) of misclassification was due to differences in the estimate of TEE for men and women.

A plot of sensitivity and specificity by rEI:BMR cut point may be used to pick the 'optimal' choice for a cut point to identify UR_{GB}. For the FFQ, the curves crossed at rEI:BMR of 1.09 for women and 1.07 for men (Figure 3). For the 24HR, the curves crossed at 1.16 for women and 1.19 for men (Figure 4).

The median expenditure was 11775 kJ and 9558 kJ for TEE_{DLW} and 11730 kJ and 8986 kJ for TEE_{GB} for men and women, respectively. The Wilcoxon signed-rank test indicated significant within-person differences between the two methods. The correlation of TEE_{GB} with TEE_{DLW} was 0.71 for men and 0.68 for women.

To better understand the implications of using the Goldberg method for classifying UR_{GB}, we compared results of relating underreporting status with personal characteristics using models previously published for UR_{DLW}. In these models, all of the odds ratio estimates for women were closer to one than in the model of UR_{DLW} (results not shown). For men, the association of BMI and the number of eating occasions were stronger in the UR_{GB} model; comparison of reported activity level with others was weaker in the model of UR_{GB}. The models of UR_{GB} on the 24HR in women showed a stronger relationship with BMI than the UR_{DLW} model (results not shown). However, the relationships with

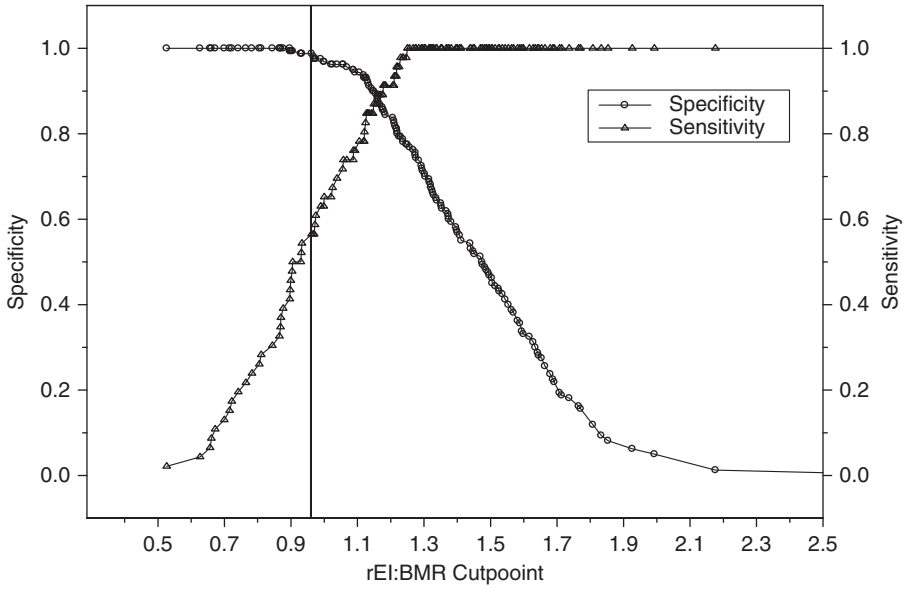

Figure 4 Plot of sensitivity and specificity by rEI:BMR cut point for the 24HR for women. The vertical line indicates the cut point from the standard Goldberg method using the values suggested by Black (2000a) (cut point = 0.96).

Table 2 Results of simulation to investigate the effect of sensitivity and specificity on estimating the relationship of a variable to underreporting status[a]

Sensitivity (%)	*Specificity (%)*	*True effect*[b]	*Estimated effect*[b]	*Relative bias (%)*	*Power using true underreporing status (%)*	*Power using Goldberg method (%)*
45	99	1.35	1.21	37.5	93	48
55	99	1.35	1.22	33.5	93	52
65	88	1.35	1.18	46.2	93	47
93	88	1.35	1.28	19.5	93	79
93	78	1.35	1.24	27.7	93	71

Abbreviation: BMI, body mass index.
[a]300 data sets of size 500 were simulated for each sensitivity/specificity combination. The predictor variable was simulated to be BMI with mean 27.9 (s.d. = 5.3).
[b]Odds ratio for 5 kg/m^2 change in BMI.

fear of negative evaluation and social desirability were not as strong in the UR_{GB} model as the UR_{DLW} model. For men on the 24HR, UR_{GB} was not as strongly related to BMI, education or restraint, as in the UR_{DLW} model. The results of the simulation study to investigate the effects using UR_{GB} rather than true UR status in studies of predicting URs indicated that low sensitivity and/or low specificity can affect both bias and power (Table 2). In particular, the parameter for the predictor variable was attenuated by 19.5–37.5%, and power was reduced from 93% to 48–79%.

Discussion

This analysis explored the utility of the Goldberg method for classifying URs on FFQ and 24HR in a large DLW study, under the assumption that the DLW analysis reflects true UR status. Overall, the Goldberg method provided excellent discrimination between URs and ARs. Sensitivity of the 24HR was similar to the estimates from other studies that used food records to assess rEI (Black, 2000b; Livingstone *et al.*, 2003), ~50% for both genders combined using the standard Goldberg method. However, it can be argued that PPV has greater utility than sensitivity for evaluating the Goldberg method. The sensitivity indicates that half of the true URs were classified as $UR_{GB.}$ This resulted in a high PPV (92%), that is, the probability that, among those who are classified as UR_{GB}, most of them really are URs. Conversely, too many URs were classified as AR_{GB}; this leads to a reduced probability that those classified as AR_{GB} really are ARs, that is, the NPV (87%) is lower than the PPV. However, the NPV is still relatively high because the prevalence of AR is approximately three-quarters of the population. In contrast, sensitivity for the FFQ was higher than for the 24HR, at 92% for the standard Goldberg method. The FFQ had the opposite tendency of the 24HR; true ARs were classified as UR_{GB}. This resulted in a PPV (88%) that was lower than the NPV (92%).

Two sources of variation that may lead to differences in classification were explored: the relationship between TEE_{DLW} and TEE_{GB}, and the cut points used, which are determined by the within-person variation in rEI and TEE. The higher sensitivity for the FFQ using the standard Goldberg method compared with the 24HR is primarily due to the differences in the assumptions about the variation in the two dietary assessment methods and subsequent calculated cut points. It is not surprising that a 24HR would have more within-person variation than an FFQ, due to day-to-day variation in intake. However, the Goldberg method makes an important assumption about the FFQ that is not made for the 24HR; it assumes that because an FFQ queries usual intake, the number of days it assesses is infinite, thereby eliminating the FFQ term for variability from the equation and tightening the cut points for the FFQ (Supplementary Material). When the actual coefficient of variation for the FFQ was used in the formula, the sensitivity dropped dramatically, to 67% overall (sensitivity still remained higher than the 24HR due to less within-person variation on the FFQ compared with the 24HR).

Differences in the estimates of TEE accounted for much of the discrepancy in classification of UR_{GB} and UR_{DLW}. We attempted to improve estimation of TEE_{GB} using PAL estimated from a physical activity questionnaire in the OPEN Study. Although this approach often led to estimates of TEE that were closer to TEE_{DLW} than TEE_{GB}, the correlation between TEE_{GB} and TEE_{DLW} was lower after adjusting for PAL. Owing to large differences that have been reported between self-report physical activity and accelerometry (Troiano *et al*., 2008) and concerns about expenditure-related bias, it is not clear that self-reported estimates of PAL provide better estimates of TEE_{GB}. However, the use of PAL estimates may be promising if objective measures of physical activity are available.

The optimal cut points for maximizing both sensitivity and specificity for the FFQ were 1.09 for women and 1.07 for men, similar to the Goldberg method cut point of 1.10. For the 24HR, the optimal cut points were 1.16 for women and 1.19 for men, which vary from the Goldberg cut point (0.96). Higher cut points may be needed to classify URs when a 24HR is used, depending on the analyst's desire to maximize sensitivity, specificity or both.

This study has limitations that warrant mention. Although DLW provides an unbiased estimate of TEE, the technique still has estimation error. Therefore, the classification of misreporters using DLW is not truly a 'gold standard'. However, the within-person variation in this study for DLW was small, so the effect of measurement error in DLW is expected to be minimal. It is also important to note that the participants in the OPEN Study were predominantly white and well-educated middle-aged adults. Their levels of URs and the association of URs with personal characteristics may differ from those in minority populations, those with lower levels of education and older or younger persons. However, although the personal characteristics identified may vary in other populations, the loss of power and effect size demonstrated in this study would be expected to occur in studies of these populations.

Another important consideration in interpreting the results of studies of underreporting is the recognition that what is termed 'underreporting' is comprised of different sources of error. By definition, underreporting represents systematic error, as opposed to day-to-day variation and other random sources of error. However, systematic error may be additive systematic error, intake-related systematic error or a combination. In a previous analysis of this data set, Kipnis *et al*. (2003) identified significant intake-related bias in the FFQ and 24HR, which comprises the underreporting error described in this manuscript. This type of error leads to bias in estimating diet–disease relationships.

Analysis of this large DLW study has demonstrated that using the Goldberg method with recommended cut points may misclassify reporting status for some individuals. When evaluating these classification measures, the question that is of most interest for a particular analysis should be considered. Analysts may want to consider choice of different cut points other than those commonly used for classifying UR status, depending on whether interest is in maximizing classification of URs or ARs. Compared with DLW, use of the Goldberg method may lead to loss of power and biased estimates of the association of UR with personal characteristics in predictive models of underreporting status, and any analysis of UR_{GB} should be interpreted in light of this. Although the sensitivity for the 24HR was low, the PPV was still high indicating that among the UR_{GB} classified by the Goldberg method most of them were true URs, and NPV was also high; these measures were also high for the FFQ. Thus, in the absence of objective measures of TEE or physical activity, the Goldberg method appears to be a reasonable technique to classify URs.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

This work was supported by a contract from the National Cancer Institute (263-MQ-612378). We thank Kristen Beavers, Sharon Kirkpatrick and Anne Rodgers for helpful suggestions on the manuscript. We would also like to acknowledge the contribution of Arthur Schatzkin to the conception and conduct of the OPEN Study, and his contributions to this manuscript.

References

Black AE (2000a). Critical evaluation of energy intake using the Goldberg cut-off for energy intake: basal metabolic rate. A practical guide to its calculation, use and limitations. *Int J Obes Relat Metab Disord* **24**, 1119–1130.

Black AE (2000b). The sensitivity and specificity of the Goldberg cut-off for EI:BMR for identifying diet reports of poor validity. *Eur J Clin Nutr* **54**, 395–404.

Black AE, Goldberg GR, Jebb SA, Livingstone MB, Cole TJ, Prentice AM (1991). Critical evaluation of energy intake data using fundamental principles of energy physiology: 2. Evaluating the results of published surveys. *Eur J Clin Nutr* **45**, 583–599.

Conway JM, Ingwersen LA, Moshfegh AJ (2004). Accuracy of dietary recall using the USDA five-step multiple-pass method in men: an observational validation study. *J Am Diet Assoc* **104**, 595–603.

Conway JM, Ingwersen LA, Vinyard BT, Moshfegh AJ (2003). Effectiveness of the US Department of Agriculture 5-step multiple-pass method in assessing food intake in obese and nonobese women. *Am J Clin Nutr* **77**, 1171–1178.

Crowne DP, Marlowe D (1960). A new scale of social desirability independent of psychopathology. *J Consult Pyschol* **24**, 349–354.

Drummond SE, Crombie NE, Cursiter MC, Kirk TR (1998). Evidence that eating frequency is inversely related to body weight status in male, but not female, non-obese adults reporting valid dietary intakes. *Int J Obes Relat Metab Disord* **22**, 105–112.

Fischer DG, Fick C (1993). Measuring social desirability-short forms of the marlowe-crowne social desirability scale. *Educ Psychol Meas* **53**, 417–424.

Goldberg GR, Black AE, Jebb SA, Cole TJ, Murgatroyd PR, Coward WA *et al* (1991). Critical evaluation of energy intake data using fundamental principles of energy physiology: 1. Derivation of cut-off limits to identify under-recording. *Eur J Clin Nutr* **45**, 569–581.

Hill RJ, Davies PS (2001). The validity of self-reported energy intake as determined using the doubly labelled water technique. *Br J Nutr* **85**, 415–430.

Hosmer DW, Lemeshow S (2000). *Applied Logistic Regression*, 2nd edn. John Wiley & Sons Inc: New York, NY, USA.

Huang TTK, Roberts SB, Howarth NC, McCrory MA (2005). Effect of screening out implausible energy intake reports on relationships between diet and BMI. *Obesity Res* **13**, 1205–1217.

Kipnis V, Subar AF, Midthune D, Freedman LS, Ballard-Barbash R, Troiano R *et al.* (2003). The structure of dietary measurement error: results of the OPEN biomarker study. *Am J Epidemiol* **158**, 14–21.

Leary MR (1983). A brief version of the fear of negative evaluation scale. *Pers Soc Psychol Bull* **9**, 371–375.

Livingstone MB, Black AE (2003). Markers of the validity of reported energy intake. *J Nutr* **133** (Suppl 3), 895S–920S.

Livingstone MB, Robson PJ, Black AE, Coward WA, Wallace JM, McKinley MC *et al* (2003). An evaluation of the sensitivity and specificity of energy expenditure measured by heart rate and the Goldberg cut-off for energy intake: basal metabolic rate for identifying mis-reporting of energy intake by adults and children: a retrospective analysis. *Eur J Clin Nutr* **57**, 455–463.

Macdiarmid J, Blundell J (1998). Assessing dietary intake: who, what and why of under-reporting. *Nutr Res Rev* **11**, 231–253.

Moshfegh AJ, Rhodes DG, Baer DJ, Murayi T, Clemens JC, Rumpler WV *et al* (2008). The US department of agriculture automated multiple-pass method reduces bias in the collection of energy intakes. *Am J Clin Nutr* **88**, 324–332.

Racette SB, Schoeller DA, Luke AH, Shay K, Hnilicka J, Kushner RF (1994). Relative dilution spaces of 2H- and 18O-labeled water in humans. *Am J Physiol* **267**, E585–E590.

Schoeller DA (1992). Isotope Dilution Methods. In: Björntorp P, Brodoff BN (eds). *Obesity*. JB Lippincott Co: New York, NY, USA, pp 80–88.

Schofield WN (1985). Predicting basal metabolic rate, new standards and review of previous work. *Hum Nutr Clin Nutr* **39** (Suppl 1), 5–41.

Strahan R, Gerbasi K (1972). Short, homogeneous versions of the Marlowe-Crowne social desirability scale. *J Clin Psychol* **28**, 191–193.

Stunkard AJ, Messick S (1985). The three-factor eating questionnaire to measure dietary restraint, disinhibition and hunger. *J Psychosom Res* **29**, 71–83.

Stunkard AJ, Sorensen T, Schulsinger F (1982). Use of the Danish Adoption Register for the study of obesity and thinness. *Res Publ Assoc Res Nerv Ment Dis* **60**, 115–120.

Subar AF, Kipnis V, Troiano RP, Midthune D, Schoeller DA, Bingham S *et al* (2003). Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: the OPEN study. *Am J Epidemiol* **158**, 1–13.

Subar AF, Thompson FE, Kipnis V, Midthune D, Hurwitz P, McNutt S *et al* (2001). Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires: the Eating at America's Table Study. *Am J Epidemiol* **154**, 1089–1099.

Tooze JA, Subar AF, Thompson FE, Troiano R, Schatzkin A, Kipnis V (2004). Psychosocial predictors of energy underreporting in a large doubly labeled water study. *Am J Clin Nutr* **79**, 795–804.

Trabulsi J, Troiano RP, Subar AF, Sharbaugh C, Kipnis V, Schatzkin A *et al* (2003). Precision of the doubly labeled water method in a large-scale application: evaluation of a streamlined-dosing protocol in the Observing Protein and Energy Nutrition (OPEN) study. *Eur J Clin Nutr* **57**, 1370–1377.

Troiano RP, Berrigan D, Dodd KW, Mâsse LC, Tilert T, McDowell M (2008). Physical activity in the United States measured by accelerometer. *Med Sci Sports Exerc* **40**, 181–188.

Supplementary Information accompanies the paper on European Journal of Clinical Nutrition website (http://www.nature.com/ejcn)